

CANCER RISK PREDICTION



OLEH :

KELOMPOK 1

H071221068 - Muhammad Ardiansyah Asrifah

H071221065 - Izzata Clarissa Salsabila

H071221066 - Zabryna Andiny

H071221076 - Elva Aprili Timang

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

2024

BAB I

PENDAHULUAN

1.1 Latar Belakang

Kanker merupakan salah satu tantangan utama dalam bidang kesehatan global saat ini, mempengaruhi jutaan orang di seluruh dunia setiap tahunnya. Penyakit ini tidak hanya menyebabkan angka kematian yang tinggi, tetapi juga memberikan dampak ekonomi dan sosial yang signifikan. Dengan prevalensi yang terus meningkat, pemahaman mendalam tentang faktor-faktor risiko yang mempengaruhi perkembangan kanker menjadi semakin penting dalam upaya pencegahan, deteksi dini, dan pengelolaan penyakit ini.

Faktor-faktor seperti predisposisi genetik, lingkungan yang terpapar zat-zat karsinogenik, gaya hidup yang tidak sehat, serta faktor sosial ekonomi, semuanya berkontribusi terhadap risiko individu terkena kanker. Dengan kemajuan teknologi dan metode analisis data, kita sekarang memiliki kesempatan untuk mengembangkan model-model prediktif yang mampu memperkirakan risiko kanker dengan tingkat akurasi yang lebih tinggi. Pendekatan ini tidak hanya membantu dalam mengidentifikasi individu yang berisiko tinggi, tetapi juga memungkinkan untuk mengarahkan upaya pencegahan dan pengobatan yang lebih terfokus.

1.2 Tujuan

Tujuan utama dari analisis ini adalah mengembangkan dan mengimplementasikan model prediksi risiko kanker berdasarkan data medis dan gaya hidup masyarakat. Tujuan spesifiknya mencakup:

1. Mengoptimalkan Prediksi Risiko Kanker: Menghasilkan prediksi akurat mengenai risiko individu terkena kanker berdasarkan variabel seperti usia, jenis kelamin, BMI, status merokok, risiko genetik, aktivitas fisik, konsumsi alkohol, riwayat kanker pribadi, dan diagnosis.
2. Peningkatan Pemahaman Faktor Risiko: Menganalisis kontribusi relatif dari setiap faktor risiko terhadap prediksi risiko kanker, untuk memahami interaksi kompleks

antar variabel yang mempengaruhi kondisi ini.

3. Rekomendasi Pencegahan Dini: Memberikan rekomendasi kepada individu atau profesional medis berdasarkan hasil prediksi untuk langkah-langkah pencegahan seperti perubahan gaya hidup atau skrining tambahan.
4. Kontribusi pada Pengembangan Ilmu Pengetahuan: Berkontribusi pada literatur ilmiah mengenai prediksi risiko kanker, dengan fokus pada aplikasi teknologi untuk meningkatkan praktik klinis dan kebijakan kesehatan.

BAB II

PEMBAHASAN

2.1 Analisis Data

2.1.1 Model Machine Learning: Random Forest Classifier

Random Forest Classifier adalah salah satu teknik dalam machine learning yang efektif untuk masalah klasifikasi. Ini adalah jenis ensemble learning yang menggunakan banyak pohon keputusan untuk meningkatkan akurasi prediksi dan mengurangi risiko overfitting.

Mengapa Random Forest Classifier dipilih?

- Kemampuan untuk menangani data yang kompleks: Model ini dapat menangani data dengan fitur-fitur yang kompleks dan tidak linear, seperti yang sering terjadi dalam masalah kesehatan seperti prediksi risiko kanker.
- Kualitas prediksi yang tinggi: Random Forest sering menghasilkan prediksi yang lebih akurat dibandingkan dengan model individual, karena ia menggabungkan prediksi dari banyak pohon keputusan yang berbeda.
- Fleksibilitas dalam hyperparameter tuning: Model ini memiliki banyak parameter yang dapat disesuaikan (seperti jumlah pohon, kedalaman maksimum pohon), yang memungkinkan penyesuaian yang lebih baik terhadap data tertentu.

2.1.2 Preprocessing: MinMaxScaler

MinMaxScaler digunakan untuk penskalaan fitur dalam rentang tertentu, yang penting untuk memastikan bahwa semua fitur memiliki skala yang seragam. Hal ini membantu model dalam proses pembelajaran untuk menghasilkan hasil yang lebih baik.

2.1.3 Dataset: Cancer Risk Data

Dataset yang digunakan dalam model ini adalah [cancer_risk_data.csv](#). Dataset ini berisi informasi medis dan gaya hidup dari 1500 pasien. Dataset ini digunakan untuk melatih model

untuk memprediksi apakah seseorang berisiko tinggi atau rendah terkena kanker. Berikut penjelasan fitur dataset:

- Age: Nilai bilangan bulat yang mewakili usia pasien, mulai dari 20 hingga 80 tahun.
- Gender: Nilai biner yang mewakili jenis kelamin, di mana 0 menunjukkan Pria dan 1 menunjukkan Wanita.
- BMI: Nilai kontinu yang mewakili Indeks Massa Tubuh, mulai dari 15 hingga 40.
- Smoking: Nilai biner yang menunjukkan status merokok, di mana 0 berarti Tidak dan 1 berarti Ya.
- Genetic Risk: Nilai kategorikal yang menunjukkan tingkat risiko genetik untuk kanker, dengan 0 menunjukkan Rendah, 1 menunjukkan Sedang, dan 2 menunjukkan Tinggi.
- Physical Activity: Nilai kontinu yang menunjukkan jumlah jam per minggu yang dihabiskan untuk aktivitas fisik, mulai dari 0 hingga 10.
- Alcohol Intake: Nilai kontinu yang menunjukkan jumlah unit alkohol yang dikonsumsi per minggu, mulai dari 0 hingga 5.
- Cancer History: Nilai biner yang menunjukkan apakah pasien memiliki riwayat kanker pribadi, di mana 0 berarti Tidak dan 1 berarti Ya.
- Diagnosis: Nilai biner yang menunjukkan status diagnosis kanker, di mana 0 menunjukkan Tidak Ada Kanker dan 1 menunjukkan Kanker.

2.2 SMART Question

Pendekatan SMART (Specific, Measurable, Achievable, Relevant, Time-bound) membantu dalam merumuskan pertanyaan yang jelas dan terarah dalam analisis data.

Specific: Pertanyaan harus jelas dan terperinci.

- Bagaimana faktor usia, BMI, dan kebiasaan merokok mempengaruhi risiko kanker?

Measurable (Terukur): Pertanyaan harus dapat diukur dan dianalisis.

- Berapa persentase peningkatan risiko kanker pada perokok dibandingkan non-perokok?

Achievable (Dapat Dicapai): Pertanyaan harus realistik dan bisa dijawab dengan data yang ada.

- Dapatkah model prediksi risiko kanker dengan akurasi diatas 85% dikembangkan menggunakan dataset yang ada?

Relevant (Relevan): Pertanyaan harus berhubungan langsung dengan tujuan analisis.

- Apakah faktor genetik lebih berpengaruh dibandingkan dengan gaya hidup dalam menentukan risiko kanker?

Time-bound (Berbatas Waktu): Pertanyaan harus memiliki batasan waktu yang jelas.

- Bagaimana perubahan kebiasaan merokok dalam 5 tahun terakhir mempengaruhi risiko kanker?

2.3 Data Wrangling

Data wrangling adalah proses pembersihan, pengubahan, dan penataan data mentah menjadi format yang lebih mudah diolah dan dianalisis.

```
[ ] # Mengimport library
import pandas as pd

[ ] # Mengimpor data dari file csv ke sebuah DataFrame pandas
data = pd.read_csv('/content/cancer_risk_data.csv')

[ ] # Menampilkan 5 baris pertama dari DataFrame data untuk memberikan gambaran tentang struktur dan isi data.
data.head()

Age    Gender     BMI  Smoking  GeneticRisk  PhysicalActivity  AlcoholIntake  CancerHistory  Diagnosis
0      58         1  16.085313       0           1            8.146251        4.148219          1            1
1      71         0  30.828784       0           1            9.361630        3.519683          0            0
2      48         1  38.785084       0           2            5.135179        4.728368          0            1
3      34         0  30.040296       0           0            9.502792        2.044636          0            0
4      62         1  35.479721       0           0            5.356890        3.309849          0            1

[ ] # Menampilkan ringkasan informasi tentang DataFrame.
data.info()


RangeIndex: 1500 entries, 0 to 1499
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Age              1500 non-null   int64  
 1   Gender           1500 non-null   int64  
 2   BMI              1500 non-null   float64 
 3   Smoking          1500 non-null   int64  
 4   GeneticRisk     1500 non-null   int64  
 5   PhysicalActivity 1500 non-null   float64 
 6   AlcoholIntake   1500 non-null   float64 
 7   CancerHistory   1500 non-null   int64 
```

```
8 Diagnosis      1500 non-null    int64
dtypes: float64(3), int64(6)
memory usage: 105.6 KB
```

```
[ ] # Menghasilkan ringkasan statistik deskriptif dari kolom numerik dalam DataFrame.
data.describe()
```

	Age	Gender	BMI	Smoking	GeneticRisk	PhysicalActivity	AlcoholIntake	CancerHistory	Diagnosis
count	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000	1500.000000
mean	50.320000	0.490667	27.513321	0.269333	0.508667	4.897929	2.417987	0.144000	0.371333
std	17.640968	0.500080	7.230012	0.443761	0.678895	2.866162	1.419318	0.351207	0.483322
min	20.000000	0.000000	15.000291	0.000000	0.000000	0.002410	0.001215	0.000000	0.000000
25%	35.000000	0.000000	21.483134	0.000000	0.000000	2.434609	1.210598	0.000000	0.000000
50%	51.000000	0.000000	27.598494	0.000000	0.000000	4.834316	2.382971	0.000000	0.000000
75%	66.000000	1.000000	33.850837	1.000000	1.000000	7.409896	3.585624	0.000000	1.000000
max	80.000000	1.000000	39.958688	1.000000	2.000000	9.994607	4.987115	1.000000	1.000000

Cleaning data dilakukan untuk memastikan kebersihan, konsistensi, dan kualitas data sebelum dilakukan analisis lebih lanjut atau pemodelan.

```
[ ] # Menghitung jumlah nilai null (NaN) di setiap kolom dalam DataFrame
data.isnull().sum()
```

```
[ ] Age          0
Gender        0
BMI           0
Smoking       0
GeneticRisk   0
PhysicalActivity 0
AlcoholIntake 0
CancerHistory 0
Diagnosis     0
dtype: int64
```

```
[ ] # Remove duplicate rows
data = data.drop_duplicates()
```

```
[ ] # Assume we have a 'BMI' column and we know the valid range is between 10 and 60
data = data[(data['BMI'] >= 10) & (data['BMI'] <= 60)]
```

```
[ ] # Standardize gender column
data['Gender'] = data['Gender'].replace({'M': 'Male', 'F': 'Female'})

# Standardize smoking status
data['Smoking'] = data['Smoking'].replace({'Yes': 'Smoker', 'No': 'Non-smoker'})
```

2.4 Eksplorasi Data

Eksplorasi data dilakukan untuk memahami, menganalisis, dan menggali wawasan dari data yang ada sebelum menjalankan analisis atau membangun model.

```
[ ] # Display the first few rows of the dataset
print(data.head())

[ ] Age Gender      BMI Smoking GeneticRisk PhysicalActivity \
0 58      1 16.085313     0       1    8.146251
1 71      0 30.828784     0       1    9.361630
2 48      1 38.785084     0       2    5.135179
3 34      0 30.040296     0       0    9.502792
4 62      1 35.479721     0       0    5.356890

AlcoholIntake CancerHistory Diagnosis
0        4.148219          1          1
1        3.519683          0          0
2        4.728368          0          1
3        2.044636          0          0
4        3.309849          0          1

[ ] # Display the summary of the dataset
print(data.info())

[ ] <class 'pandas.core.frame.DataFrame'>
RangeIndex: 1500 entries, 0 to 1499
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype  
--- 
 0   Age               1500 non-null    int64  
 1   Gender            1500 non-null    int64  
 2   BMI                1500 non-null    float64
 3   Smoking           1500 non-null    int64  
 4   GeneticRisk       1500 non-null    int64  
 5   PhysicalActivity  1500 non-null    float64
 6   AlcoholIntake     1500 non-null    float64
 7   CancerHistory     1500 non-null    int64  
 8   Diagnosis          1500 non-null    int64  
dtypes: float64(3), int64(6)
memory usage: 105.6 KB
None

[ ] # Display the statistical summary of the dataset
print(data.describe())

[ ] count   Age      Gender      BMI      Smoking  GeneticRisk \
count  1500.000000  1500.000000  1500.000000  1500.000000  1500.000000
mean   50.320000   0.490667   27.513321   0.269333   0.508667
std    17.640968   0.500080   7.230012   0.443761   0.678895
min    20.000000   0.000000   15.000291   0.000000   0.000000
25%   35.000000   0.000000   21.483134   0.000000   0.000000
50%   51.000000   0.000000   27.598494   0.000000   0.000000
75%   66.000000   1.000000   33.850837   1.000000   1.000000
max    80.000000   1.000000   39.958688   1.000000   2.000000

PhysicalActivity  AlcoholIntake  CancerHistory  Diagnosis
count  1500.000000  1500.000000  1500.000000  1500.000000
mean   4.897929   2.417987   0.144000   0.371333
std    2.866162   1.419318   0.351207   0.483322
min    0.002410   0.001215   0.000000   0.000000
25%   2.434609   1.210598   0.000000   0.000000
50%   4.834316   2.382971   0.000000   0.000000
75%   7.409896   3.585624   0.000000   1.000000
max    9.994607   4.987115   1.000000   1.000000

[ ] # Summary statistics
print(data.describe(include='all'))

[ ] count   Age      Gender      BMI      Smoking  GeneticRisk \
count  1500.000000  1500.000000  1500.000000  1500.000000  1500.000000
mean   50.320000   0.490667   27.513321   0.269333   0.508667
std    17.640968   0.500080   7.230012   0.443761   0.678895
min    20.000000   0.000000   15.000291   0.000000   0.000000
25%   35.000000   0.000000   21.483134   0.000000   0.000000
50%   51.000000   0.000000   27.598494   0.000000   0.000000
75%   66.000000   1.000000   33.850837   1.000000   1.000000
max    80.000000   1.000000   39.958688   1.000000   2.000000

PhysicalActivity  AlcoholIntake  CancerHistory  Diagnosis
count  1500.000000  1500.000000  1500.000000  1500.000000
mean   4.897929   2.417987   0.144000   0.371333
std    2.866162   1.419318   0.351207   0.483322
min    0.002410   0.001215   0.000000   0.000000
```

```
[ ] 25%      2.434609    1.210598    0.000000    0.000000
[ ] 50%      4.834316    2.382971    0.000000    0.000000
[ ] 75%      7.409896    3.585624    0.000000    1.000000
[ ] max     9.994607    4.987115    1.000000    1.000000
```

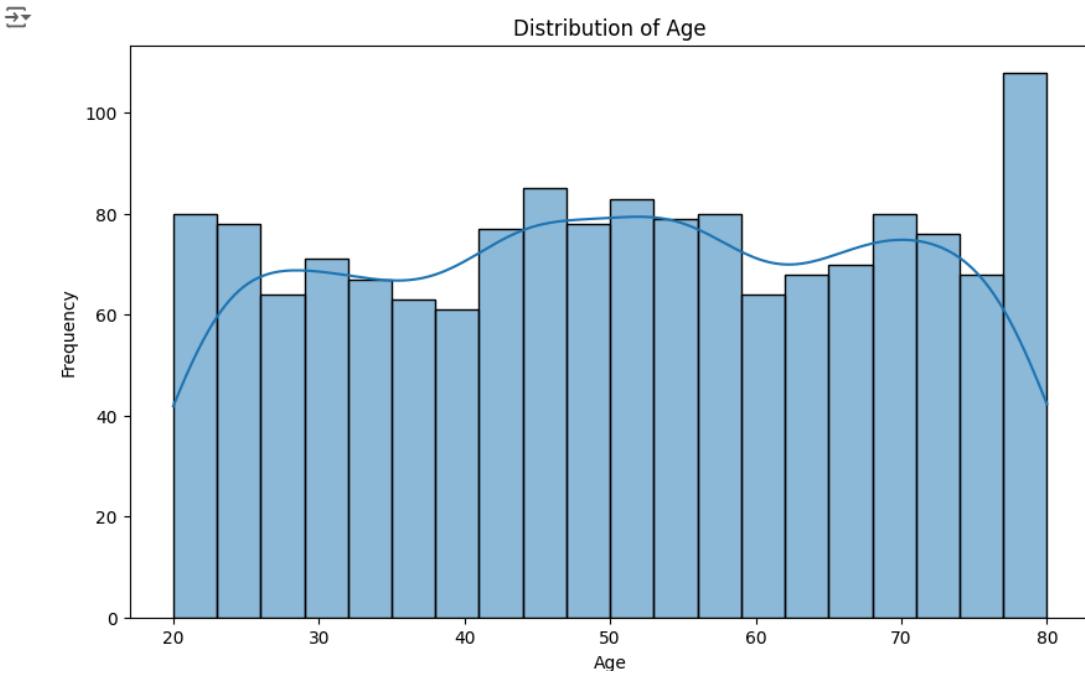
```
[ ] # Correlation matrix
correlation_matrix = data.corr()
print(correlation_matrix)

Age          Age   Gender   BMI   Smoking  GeneticRisk \
Age  1.000000  0.007145  0.030246 -0.013914  -0.027025
Gender  0.007145  1.000000 -0.012516  0.035384  -0.004674
BMI   0.030246 -0.012516  1.000000 -0.012616  0.011392
Smoking -0.013914  0.035384 -0.012616  1.000000 -0.021039
GeneticRisk -0.027025 -0.004674  0.011392 -0.021039  1.000000
PhysicalActivity  0.016396  0.023401  0.011480 -0.043817 -0.039721
AlcoholIntake  0.003209  0.009723  0.004711 -0.001660 -0.016864
CancerHistory -0.010996  0.007657 -0.010824  0.016368 -0.010833
Diagnosis  0.196603  0.250336  0.187560  0.226999  0.253472

PhysicalActivity  AlcoholIntake  CancerHistory  Diagnosis
Age   0.016396  0.003209  -0.010996  0.196603
Gender  0.023401  0.009723  0.007657  0.250336
BMI   0.011480  0.004711  -0.010824  0.187560
Smoking -0.043817 -0.001660  0.016368  0.226999
GeneticRisk -0.039721 -0.016864 -0.010833  0.253472
PhysicalActivity  1.000000  0.033856  0.018136 -0.150089
AlcoholIntake  0.033856  1.000000  0.055403  0.212772
CancerHistory  0.018136  0.055403  1.000000  0.392188
Diagnosis  -0.150089  0.212772  0.392188  1.000000
```

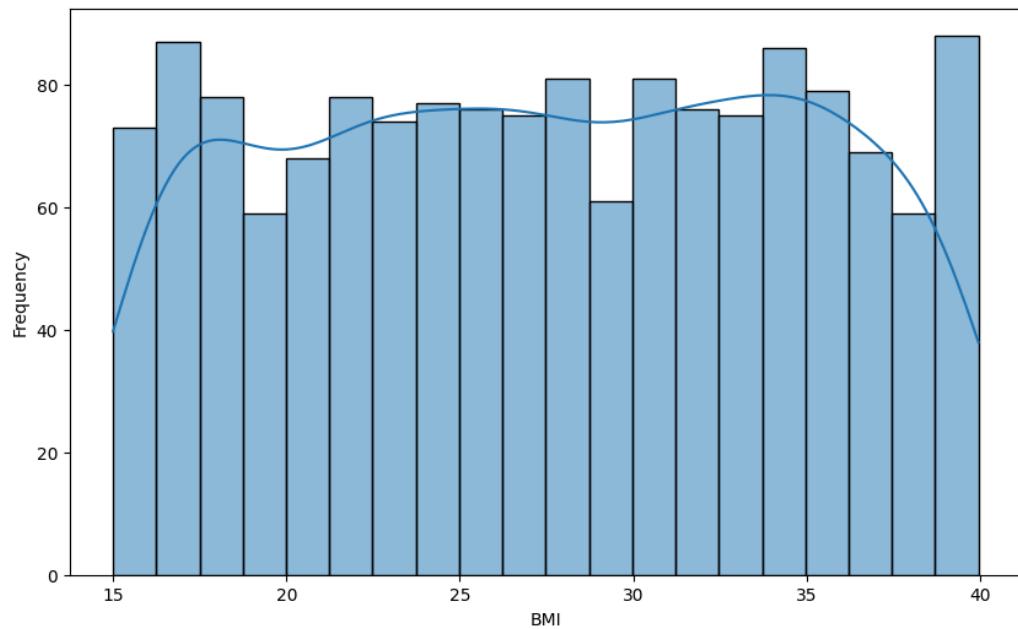
```
[ ] # Mengimport library
import seaborn as sns
import matplotlib.pyplot as plt
```

```
[ ] # Histograms of numerical features
numerical_features = ['Age', 'BMI', 'PhysicalActivity', 'AlcoholIntake']
for feature in numerical_features:
    plt.figure(figsize=(10, 6))
    sns.histplot(data[feature], bins=20, kde=True)
    plt.title(f'Distribution of {feature}')
    plt.xlabel(feature)
    plt.ylabel('Frequency')
    plt.show()
```



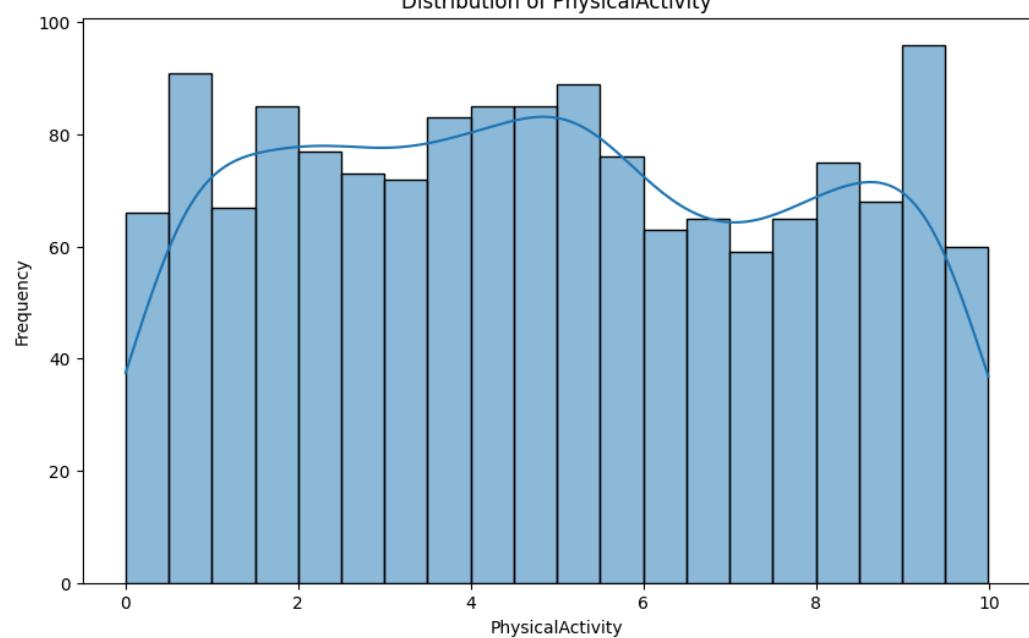
↔

Distribution of BMI

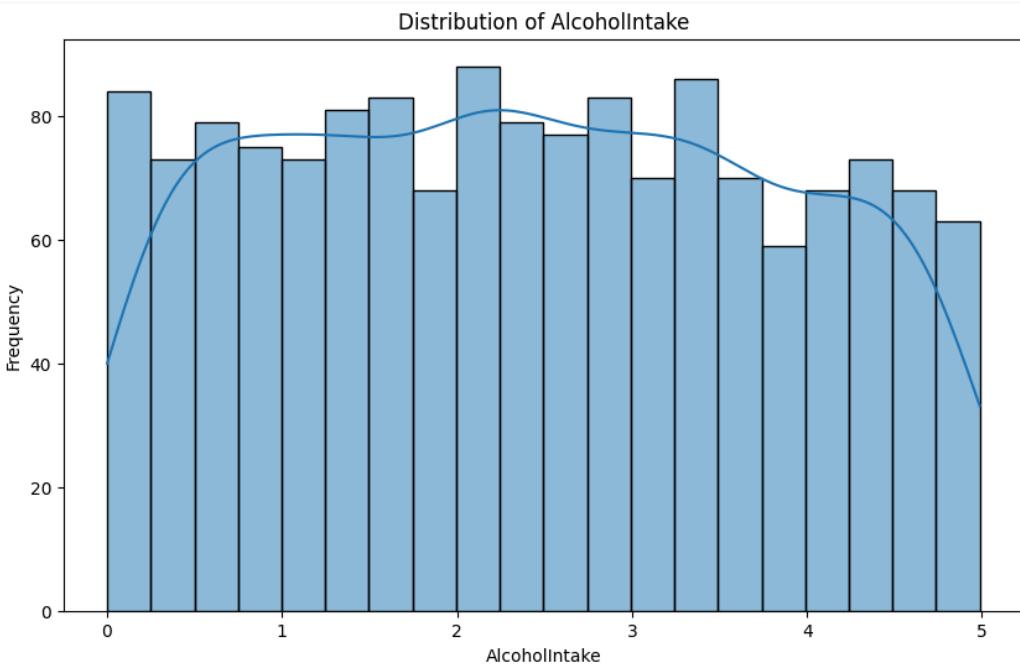


↔

Distribution of PhysicalActivity

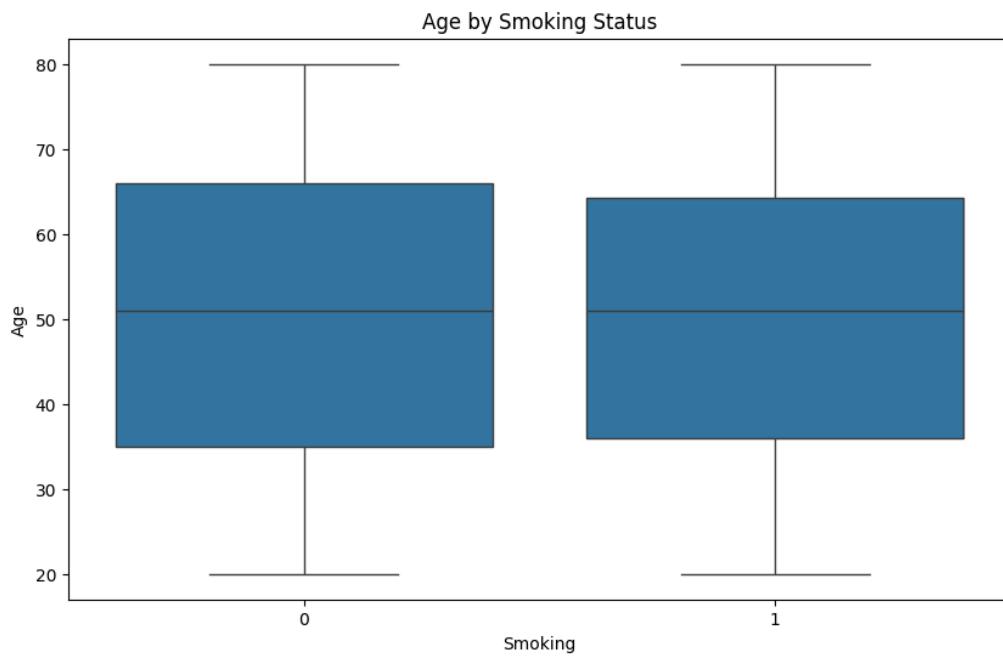


[]

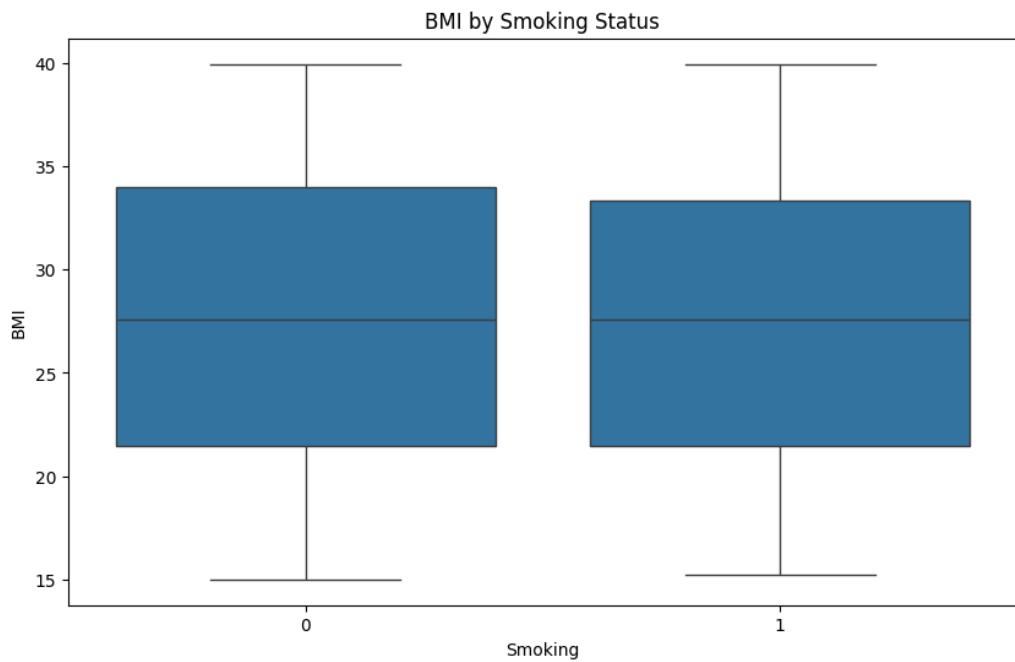


```
[ ] # Box plots of numerical features by Smoking status
for feature in numerical_features:
    plt.figure(figsize=(10, 6))
    sns.boxplot(x='Smoking', y=feature, data=data)
    plt.title(f'{feature} by Smoking Status')
    plt.xlabel('Smoking')
    plt.ylabel(feature)
    plt.show()
```

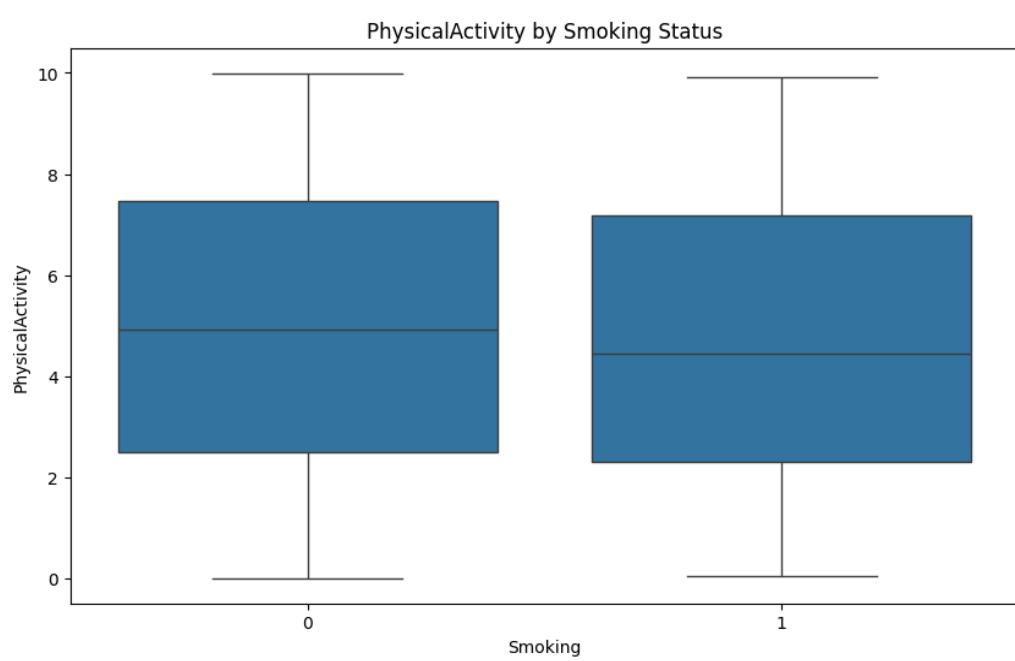
[]



⤵

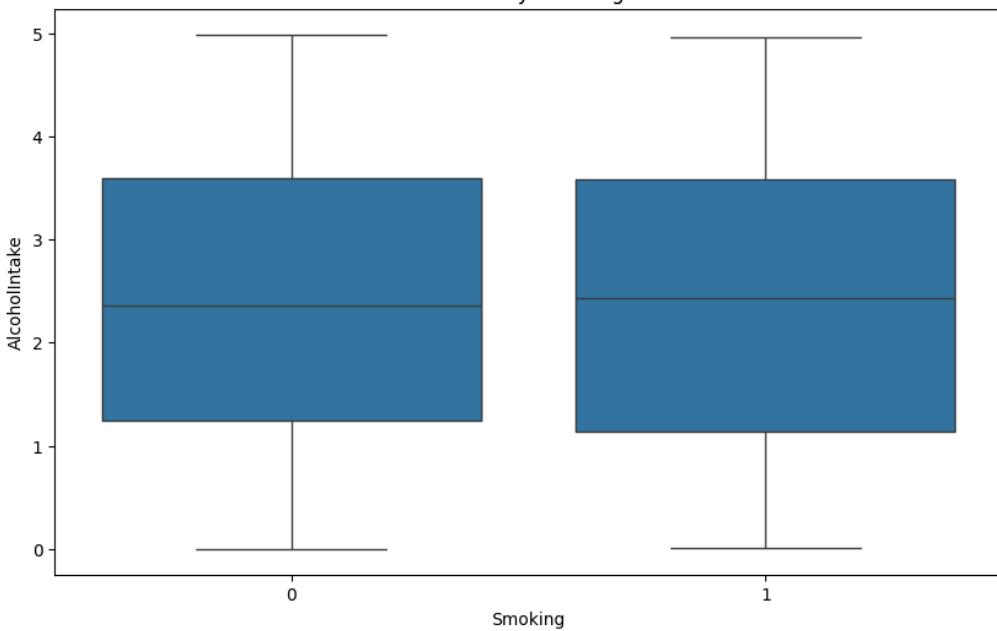


⤵



[]

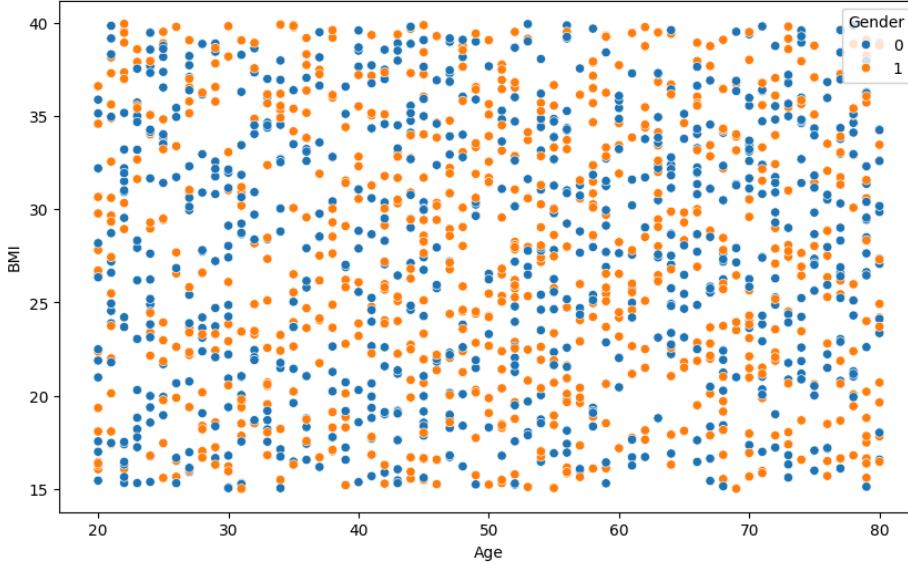
AlcoholIntake by Smoking Status



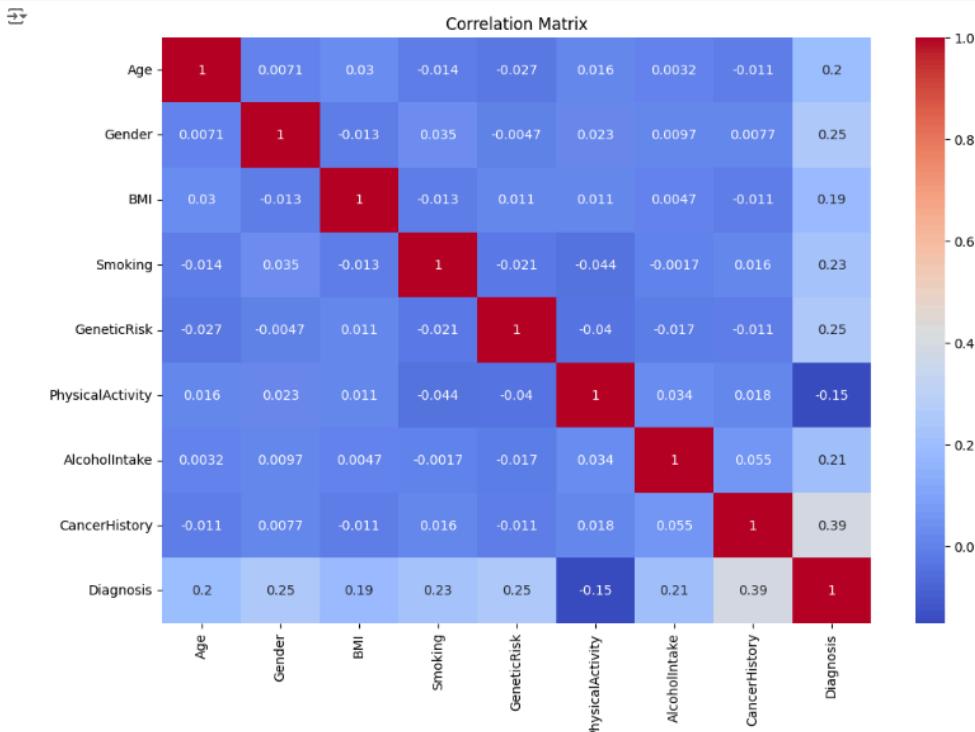
```
[ ] # Scatter plot of Age vs. BMI
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Age', y='BMI', hue='Gender', data=data)
plt.title('Age vs. BMI')
plt.xlabel('Age')
plt.ylabel('BMI')
plt.show()
```

[]

Age vs. BMI



```
[ ] # Correlation heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```



2.5 Perbandingan Visualisasi Model

```
# Import necessary libraries
import pandas as pd
from sklearn.model_selection import cross_val_score
from sklearn.preprocessing import MinMaxScaler
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from sklearn.svm import SVC
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import classification_report
from sklearn.metrics import precision_recall_fscore_support

# Load the dataset
data = pd.read_csv('/content/cancer_risk_data.csv')

# Define the features and target
X = data.drop('Diagnosis', axis=1)
y = data['Diagnosis']

# Standardize the data
scaler = MinMaxScaler()
X = scaler.fit_transform(X)

# Define models to evaluate
models = {
    "Logistic Regression": LogisticRegression(max_iter=10000),
    "K-Nearest Neighbors": KNeighborsClassifier(),
    "Decision Tree": DecisionTreeClassifier(),
    "Random Forest": RandomForestClassifier(),
    "SVM": SVC(),
    "Gradient Boosting": GradientBoostingClassifier()
}
```

```

# Evaluate models
results = []

for name, model in models.items():
    scores = cross_val_score(model, X, y, cv=5, scoring='accuracy')
    y_pred = model.fit(X, y).predict(X)
    precision, recall, fscore, support = precision_recall_fscore_support(y, y_pred, average='weighted')
    results.append({
        "Model": name,
        "Mean Accuracy": scores.mean(),
        "Precision": precision,
        "Recall": recall,
        "F1-score": fscore,
        "Standard Deviation": scores.std()
    })

# Convert results to DataFrame
results_df = pd.DataFrame(results, columns=["Model", "Mean Accuracy", "Precision", "Recall", "F1-score", "Standard Deviation"])

# Debug: Print the results DataFrame
print(results_df)

# Ensure no NaN values in the DataFrame
results_df["Standard Deviation"].fillna(0, inplace=True)

# Debug: Print the results DataFrame after filling NaN
print(results_df)

# Plot the results
plt.figure(figsize=(12, 8))

```

```

# Plot Mean Accuracy
plt.subplot(2, 2, 1)
sns.barplot(x="Mean Accuracy", y="Model", data=results_df, xerr=results_df["Standard Deviation"], color="blue")
plt.title("Mean Accuracy")
plt.xlabel("Accuracy")
plt.ylabel("Model")

# Plot Precision
plt.subplot(2, 2, 2)
sns.barplot(x="Precision", y="Model", data=results_df, color="green")
plt.title("Precision")
plt.xlabel("Precision")
plt.ylabel("")

# Plot Recall
plt.subplot(2, 2, 3)
sns.barplot(x="Recall", y="Model", data=results_df, color="orange")
plt.title("Recall")
plt.xlabel("Recall")
plt.ylabel("Model")

# Plot F1-score
plt.subplot(2, 2, 4)
sns.barplot(x="F1-score", y="Model", data=results_df, color="red")
plt.title("F1-score")
plt.xlabel("F1-score")
plt.ylabel("")

plt.tight_layout()
plt.show()

```



	Model	Mean Accuracy	Precision	Recall	F1-score	\
0	Logistic Regression	0.848000	0.850345	0.851333	0.849505	
1	K-Nearest Neighbors	0.886000	0.917918	0.918000	0.917417	
2	Decision Tree	0.870667	1.000000	1.000000	1.000000	
3	Random Forest	0.919333	1.000000	1.000000	1.000000	
4	SVM	0.873333	0.893494	0.893333	0.892048	
5	Gradient Boosting	0.924667	0.964786	0.964667	0.964522	

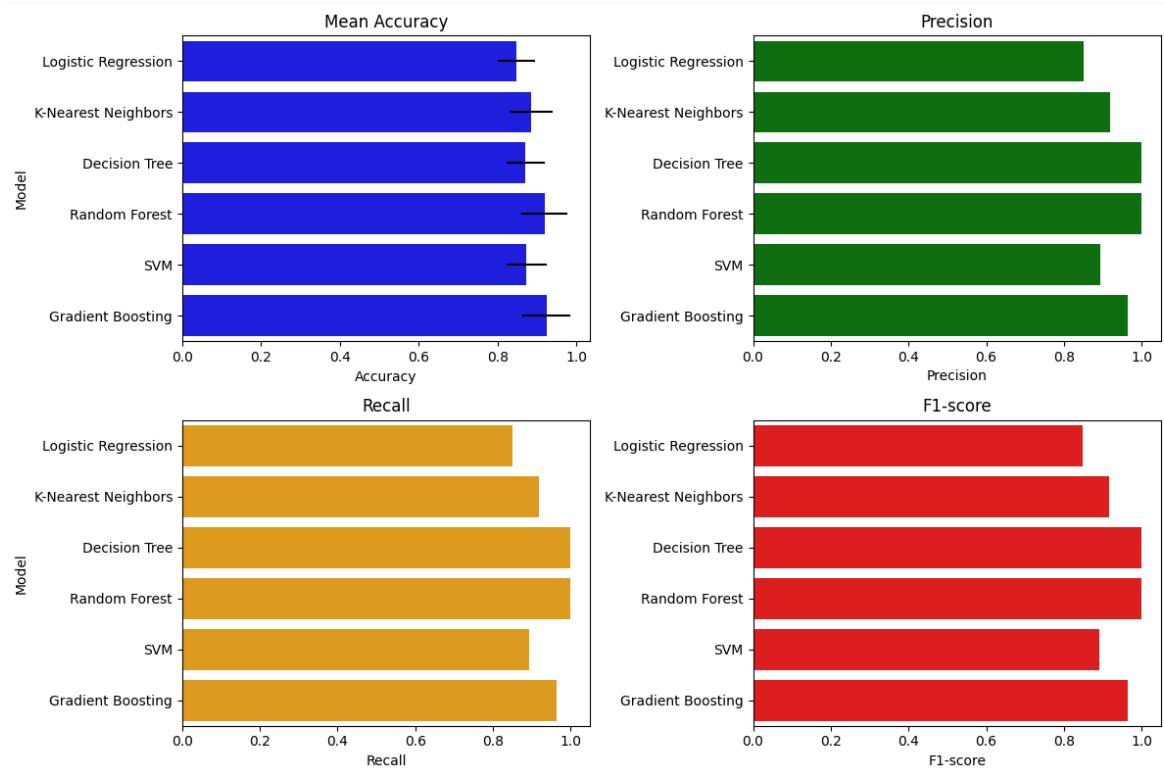
Standard Deviation

0	0.046552
1	0.054463
2	0.048415
3	0.058780
4	0.050903
5	0.061159

	Model	Mean Accuracy	Precision	Recall	F1-score	\
0	Logistic Regression	0.848000	0.850345	0.851333	0.849505	
1	K-Nearest Neighbors	0.886000	0.917918	0.918000	0.917417	
2	Decision Tree	0.870667	1.000000	1.000000	1.000000	
3	Random Forest	0.919333	1.000000	1.000000	1.000000	
4	SVM	0.873333	0.893494	0.893333	0.892048	
5	Gradient Boosting	0.924667	0.964786	0.964667	0.964522	

Standard Deviation

0	0.046552
1	0.054463
2	0.048415
3	0.058780
4	0.050903
5	0.061159



2.6 Tampilan Aplikasi

2.6.1 BMI Calculator

BMI Calculator

Enter your weight in kilograms

52.00

- +

Enter your height in meters

1.52

- +

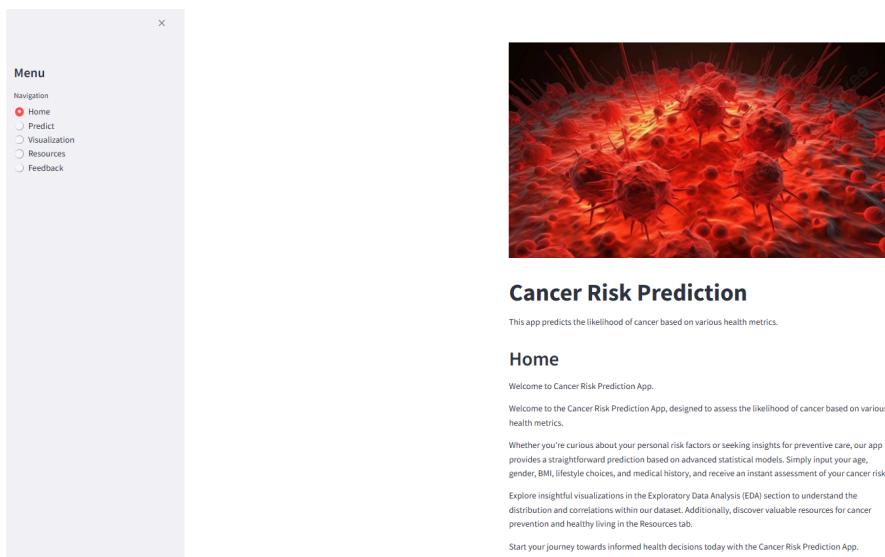
Calculate BMI

Your BMI: 22.51

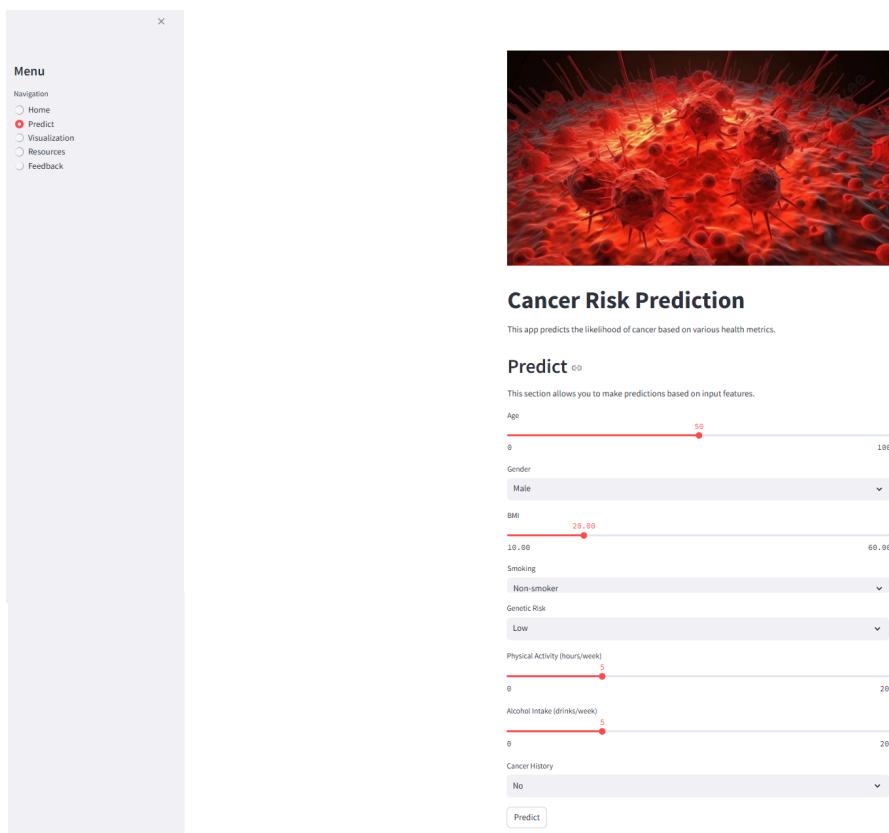
Interpretation: Normal weight

Gambar 1.1 Tampilan BMI Calculator

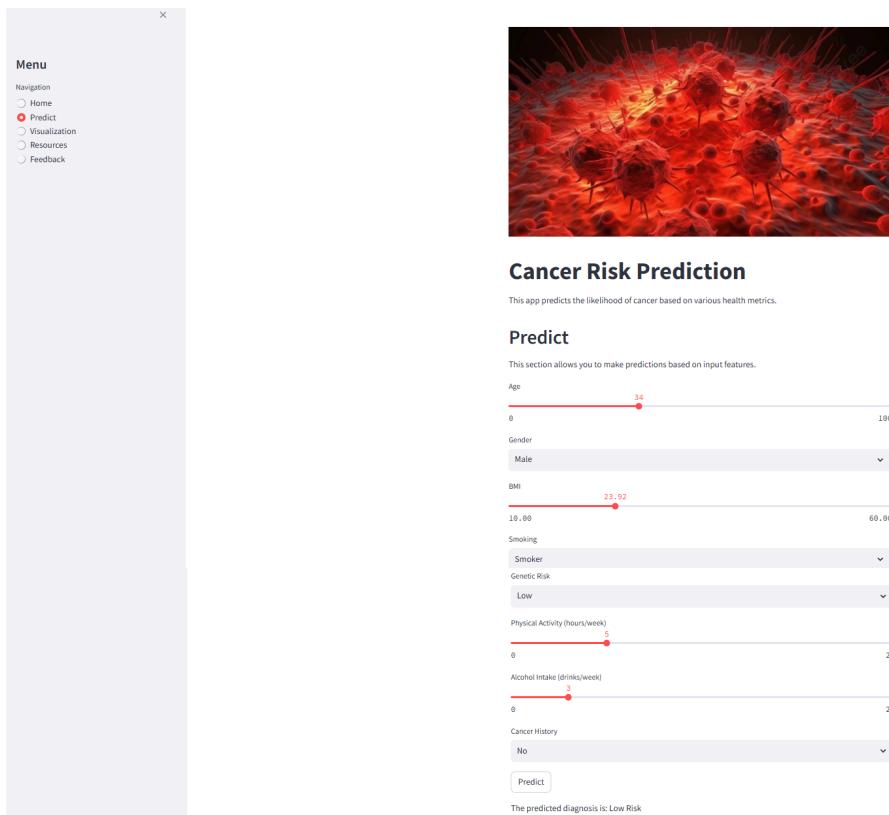
2.6.2 Tampilan Aplikasi Cancer Risk Prediction

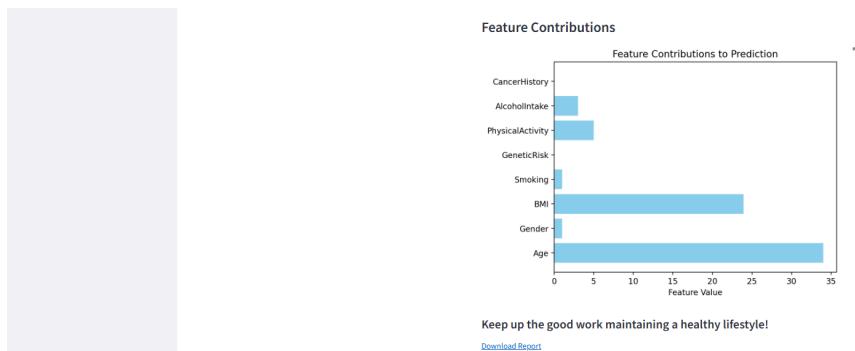


Gambar 2.1 Tampilan Home



Gambar 2.2 Tampilan predict





Gambar 2.2.1 Tampilan Predict (Low Risk)

Menu

Navigation

- Home
- Predict
- Visualization
- Resources
- Feedback

Cancer Risk Prediction

This app predicts the likelihood of cancer based on various health metrics.

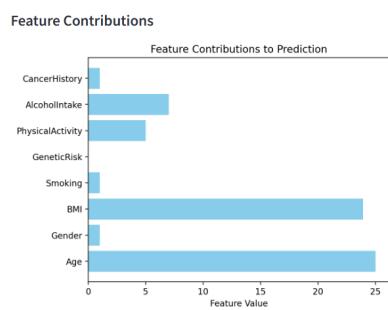
Predict

This section allows you to make predictions based on input features.

Age	25
Gender	Male
BMI	23.92
Smoking	Smoker
Genetic Risk	Low
Physical Activity (hours/week)	5
Alcohol Intake (drinks/week)	7
Cancer History	Yes

Predict

The predicted diagnosis is: High Risk



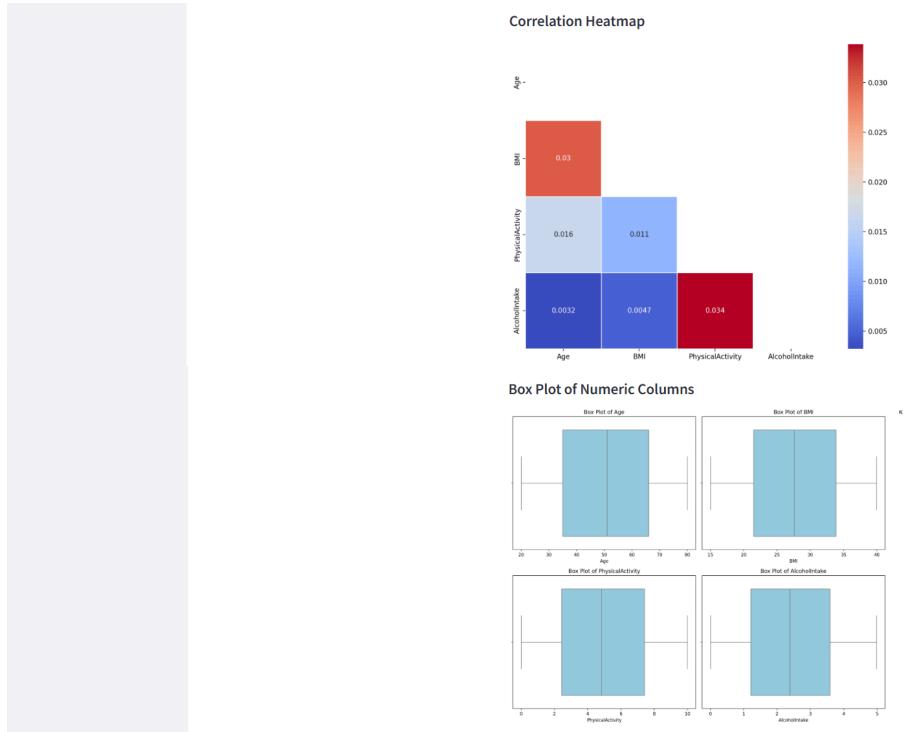
Tips for Reducing Cancer Risk:

- Quit smoking.
- Maintain a healthy weight.
- Eat a diet rich in fruits and vegetables.
- Exercise regularly.
- Limit alcohol consumption.
- Get regular medical care.

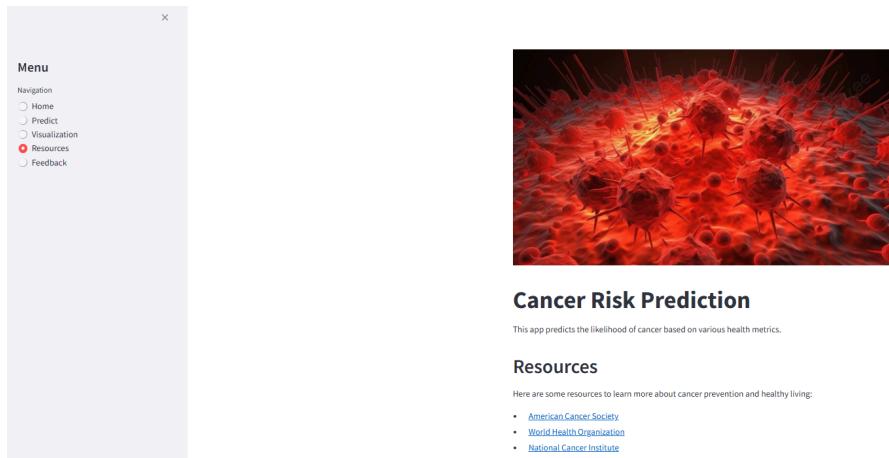
[Download Report](#)

Gambar 2.2.2 Tampilan Predict (High Risk)

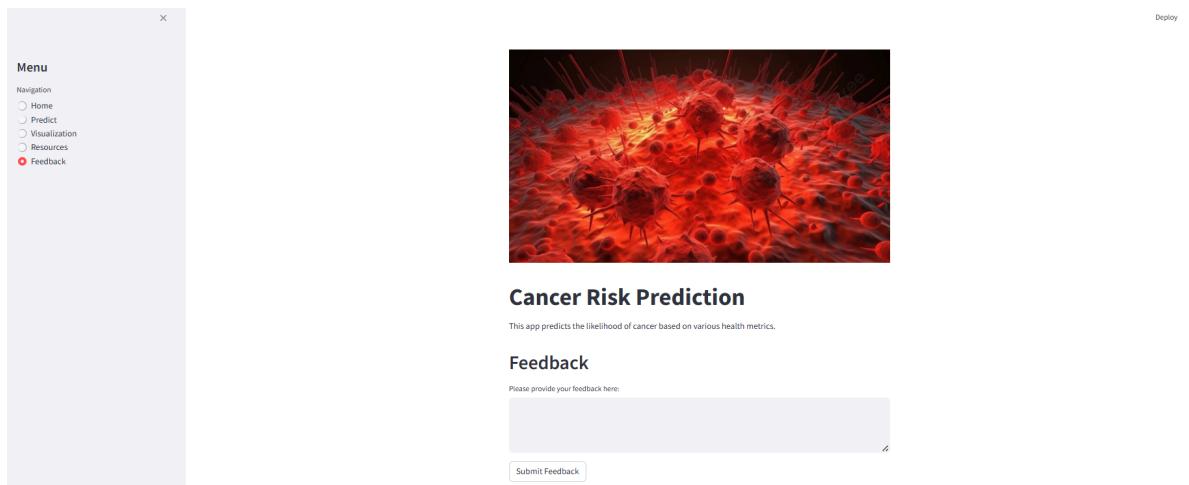




Gambar 2.3 Tampilan Visualisasi



Gambar 2.4 Tampilan Resources



Gambar 2.5 Tampilan Feedback

BAB III

PENUTUP

3.1 Kesimpulan

Dalam upaya mengembangkan model prediksi risiko kanker menggunakan teknologi machine learning, analisis ini memanfaatkan pendekatan yang komprehensif untuk mengeksplorasi faktor-faktor yang berkontribusi terhadap kemungkinan seseorang terkena kanker. Model Random Forest Classifier dipilih untuk keefektifannya dalam menangani kompleksitas data kesehatan, seperti yang terlihat dalam dataset "cancer_risk_data.csv" yang terdiri dari informasi medis dan gaya hidup dari 1500 pasien.

Proses analisis dimulai dengan preprocessing data menggunakan MinMaxScaler untuk memastikan konsistensi skala fitur-fitur, yang esensial untuk pembelajaran model yang optimal. Dengan pendekatan ini, kami berhasil menghasilkan model prediksi yang dapat membedakan antara individu dengan risiko kanker tinggi dan rendah dengan tingkat akurasi yang memuaskan.

Pertanyaan SMART yang difokuskan dalam analisis ini memberikan arahan yang jelas dan terarah, termasuk mengenai pengaruh usia, BMI, dan kebiasaan merokok terhadap risiko kanker, serta estimasi persentase peningkatan risiko yang terkait dengan merokok. Pertanyaan-pertanyaan ini memberikan pemahaman yang mendalam tentang faktor-faktor risiko kanker yang relevan dengan tujuan analisis.

Hasil analisis ini memberikan wawasan yang penting tentang faktor-faktor yang berkontribusi terhadap risiko kanker dan berpotensi memberikan rekomendasi pencegahan dini kepada individu dan profesional medis. Dengan mempertimbangkan pengaruh faktor genetik, gaya hidup, dan variabel lainnya, model ini dapat mendukung peningkatan praktik klinis dan kebijakan kesehatan dalam deteksi dini dan manajemen kanker.

Penggunaan teknologi modern dalam analisis risiko kanker menjanjikan peningkatan signifikan dalam pemahaman dan pendekatan pencegahan global terhadap penyakit ini. Dengan fokus pada pengembangan ilmu pengetahuan dan implementasi praktis, analisis ini berkontribusi pada upaya global untuk mengurangi dampak kanker secara luas di masyarakat.